Digital Quantification of Chemical Oligonucleotide Synthesis Errors

Stefan Filges ,[a] Pia Mouhanna,[a] and Anders Ståhlberg [a,b,c,*]

BACKGROUND: Chemically synthesized oligonucleotides are vital to most nucleic acids-based technologies and several applications are sensitive to oligonucleotide sequence errors. However, it is challenging to identify and quantify the types and amount of errors in synthetic oligonucleotides.

METHODS: We applied a digital sequencing approach using unique molecular identifiers to quantify errors in chemically synthesized oligonucleotides from multiple manufacturers with different synthesis strategies, purity grades, batches, and sequence context.

RESULTS: We detected both deletions and substitutions in chemical oligonucleotide synthesis, but deletions were 7 times more common. We found that 97.2% of all analyzed oligonucleotide molecules were intact across all manufacturers and purity grades, although the number of oligonucleotide molecules with deletions ranged between 0.2% and 11.7% for different types. Different batches of otherwise identical oligonucleotide types also varied significantly, and batch effect can impact oligonucleotide quality more than purification. We observed a bias of increased deletion rates in chemically synthesized oligonucleotides toward the 5'-end for 1 out of 2 sequence configurations. We also demonstrated that the performance of sequencing assays depends on oligonucleotide quality.

CONCLUSIONS: Our data demonstrate that manufacturer, synthesis strategy, purity, batch, and sequence context all contribute to errors in chemically synthesized oligonucleotides and need to be considered when choosing and evaluating oligonucleotides. High-performance oligonucleotides are essential in numerous molecular applications, including clinical diagnostics.

Introduction

Synthetic oligonucleotides are critical elements in almost all molecular biology methods. They are used, among others, for PCR (1), reverse transcription (2), sequencing (3), gene editing (4), synthetic biology (5), cloning (6), single nucleotide polymorphism assays (7), microarrays (8), electrophoretic mobility shift assays (9), and small RNA research (10). Most of these technical advances require synthetic oligonucleotides with high-quality, integrity, and purity (11).

Chemical synthesis of oligonucleotides remained cumbersome and labor intensive until the contemporary oligonucleotide synthesis method based on phosphoramidite monomers was developed in 1983 (12). In phosphoramidite synthesis, the first nucleotide is bound to a glass bead at its 3' carbon to which subsequent nucleotides are attached. Thus synthesis proceeds from 3' to 5', the opposite direction compared to enzymatic nucleic acid synthesis (13). The coupling efficiency of each attached nucleotide is generally between 98.5% and 99.5% (14, 15). While seemingly high, the consequence of a coupling efficiency of 98.5% is that for a 140 bp-long oligonucleotide 90% of all synthesized molecules will be truncated. If the coupling efficiency is increased to 99.5% efficiency, 50% of all molecules will have full length. In case of coupling failure, the reactive 5' hydroxyl group is capped by acetylation to prevent further additions of nucleotides that would result in missing bases. Incomplete oligonucleotides can be removed using different methods, such as PAGE and high-performance liquid chromatography, which can remove most oligonucleotides containing 2 or more missing nucleotides (16). However, a fraction of all truncated molecules remains after purification, especially

[a]Department of Laboratory Medicine, Sahlgrenska Center for Cancer Research, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden; [b]Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Sweden; [c]Department of Clinical Genetics and Genomics, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden.

*Address correspondence to this author at: Sahlgrenska Center for Cancer Research, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Medicinaregatan 1F 405 30 Gothenberg, Sweden. E-mail: anders.stahlberg@gu.se.

oligonucleotides with a single missing base. Other methods to remove synthesis errors include hybridization selection (17) and sequencing-based retrieval (18), which are expensive methods that require labor-intensive workflows and specialized equipment. The effects of using oligonucleotides with erroneous sequences depend on method and downstream application, but with an emerging need for single-molecule resolution, the dependency on correctly synthesized oligonucleotides also increases. Traditional quality controls of oligonucleotide synthesis, such as mass spectrometry, have improved over time but are unable to resolve the actual oligonucleotide sequence. Few studies of synthetic oligonucleotide errors exist (19, 20), and these have typically focused on describing the errors in short oligonucleotides, typical for PCR primers. In contrast, several recently developed methods depend on longer and more complex oligonucleotides, such as those containing randomized sequences and secondary structures.

Here, we applied ultrasensitive sequencing using unique molecular identifiers (UMIs) to remove polymerase-induced errors, enabling digital sequencing and detailed studies of chemically synthesized oligonucleotides. We investigated the error profiles of 4 different manufacturers, several purity grades, 2 different oligonucleotide sequences, and up to 3 different batches of the same oligonucleotide type. We also analyzed the performance of different hairpin barcoding assays using different oligonucleotide types. By profiling several types of chemically synthesized oligonucleotides using digital sequencing, we provide quantitative information about synthesis errors that is potentially critical when developing and validating nucleic acids-based methods and applications.

Materials and Methods

DNA SOURCES

Chemically synthesized oligonucleotides with 2 alternative sequences (Fig. 1, A, online Supplemental Table 1) were purchased from 4 different manufacturers with varying purity grades as listed in Table 1. For a subset of oligonucleotide types, 2 additional batches were ordered with at least 2 weeks between orders. Oligonucleotides were reconstituted in Tris-EDTA buffer, consisting of 10 mmol/L Tris-HCl and 1 mmol/L EDTA adjusted to pH 8.0 and diluted to 10^4 molecules per µL. Diluted oligonucleotides were stored at $-20\,°C$. Cell line control DNA was prepared from MCF-7 cells as described in the online Supplemental Materials and Methods.

SIMSEN-SEQ LIBRARY CONSTRUCTION

Primers were designed using National Center for Biotechnology Information Primer-Blast (21) as previously described (22). The target primers used were

validated using quantitative PCR and fragment analysis to ensure efficiency and specificity (Supplemental Fig. 1 and Supplemental Materials and Methods). SiMSen-seq (simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection using next-generation sequencing) were constructed as described in the Supplemental Materials and Methods and as described previously (22, 23). Briefly, template molecules were barcoded in triplicate 10 µL reactions, containing 0.1 U Phusion high fidelity polymerase, $1\times$ high fidelity buffer, 200 µmol/L deoxynucleotide triphosphate (all Thermo Fisher Scientific), 0.5 mol/L L-carnitine (Sigma-Aldrich), and 40 nmol/L of each SiMSen-Seq barcoding primer (Integrated DNA Technologies [IDT], PAGE-purified; online Supplemental Table 2) and 2 µL either diluted synthetic oligonucleotides corresponding to 20 000 single-stranded molecules or 50 ng genomic MCF-7 DNA. After completion of the barcoding PCR, 20 µmol/L of Tris-EDTA buffer containing 30 µg/mL protease from *Streptomyces griseus* (Sigma-Aldrich) was added to terminate each reaction. Next, a 40 µL adapter PCR was performed, containing $1\times$ Q5 Hot Start High-Fidelity Mastermix (New England BioLabs) and 400 nmol/L of each Illumina adapter primer (Sigma-Aldrich, desalted; online Supplemental Tables 3 and 4) and 10 µL diluted barcoded PCR product. Specific PCR products were purified using the Agencourt AMPure XP system (Beckman Coulter Diagnostics) according to the manufacturer's instructions. Prior to sequencing, library integrity and purity were assessed on a fragment analyzer using the NGS HS kit (both Agilent) according to the manufacturer's instructions.

SEQUENCING AND DATA ANALYSIS

Individual libraries were equimolarly pooled and quantified using the Qubit dsDNA HS kit (Thermo Fisher Scientific). Sequencing was performed on an Illumina MiniSeq instrument in single-end 150 bp mode according to the manufacturer instructions. Raw sequencing data were analyzed using a modified version of Debarcer as previously described (22). Briefly, errors in the sequencing data were corrected by grouping reads into barcode families for barcodes that were observed at least 10 times and forming an error-corrected consensus sequence. For additional details about sequencing, data analysis, and statistics, see Supplemental Material and Methods.

DATA AVAILABILITY

Raw sequencing data were deposited in the National Center for Biotechnology Information Sequence Read Archive (SRA, BioProject PRJNA727098).
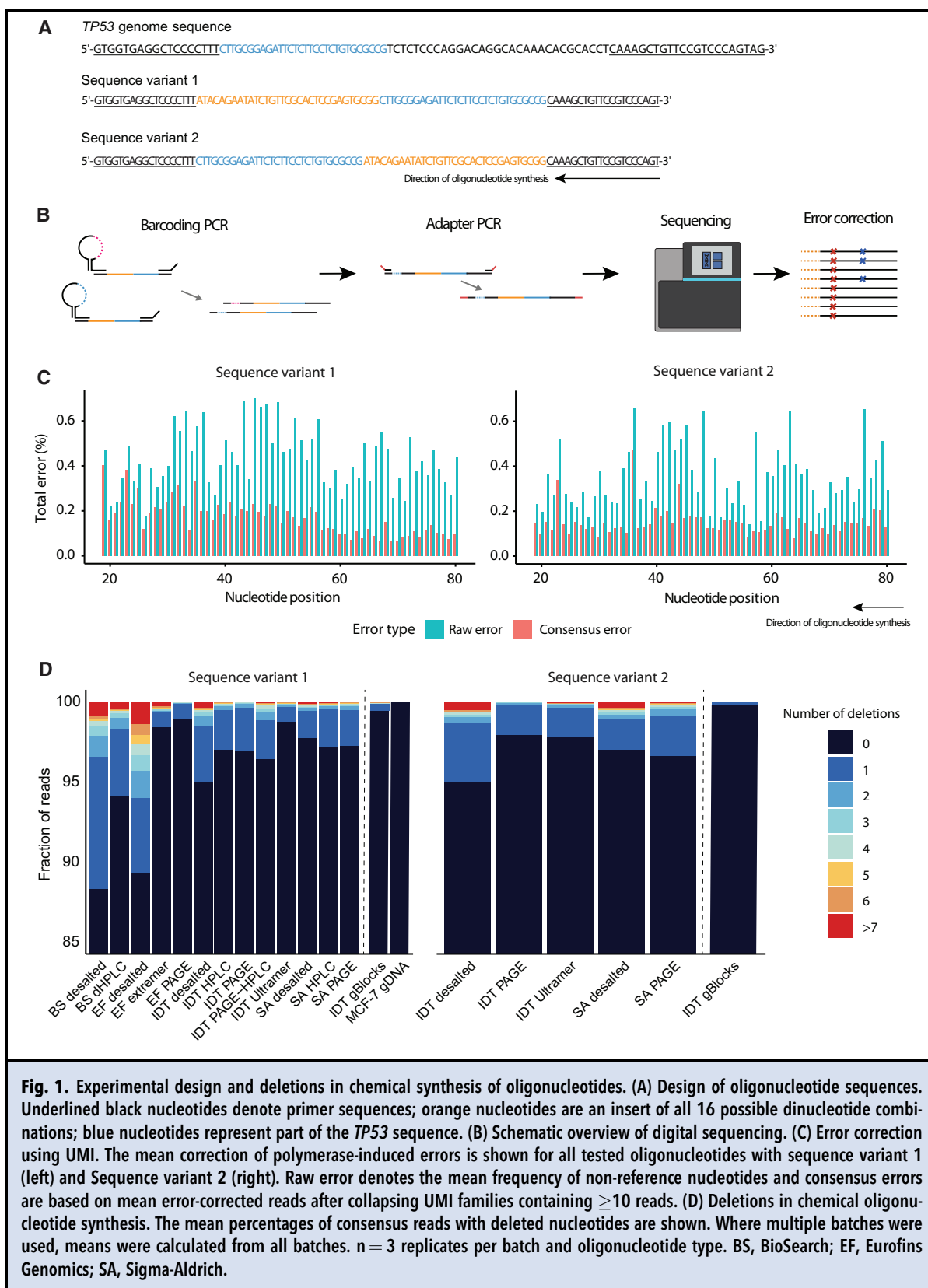
Fig. 1. Experimental design and deletions in chemical synthesis of oligonucleotides. (A) Design of oligonucleotide sequences. Underlined black nucleotides denote primer sequences; orange nucleotides are an insert of all 16 possible dinucleotide combinations; blue nucleotides represent part of the *TP53* sequence. (B) Schematic overview of digital sequencing. (C) Error correction using UMI. The mean correction of polymerase-induced errors is shown for all tested oligonucleotides with sequence variant 1 (left) and Sequence variant 2 (right). Raw error denotes the mean frequency of non-reference nucleotides and consensus errors are based on mean error-corrected reads after collapsing UMI families containing ≥10 reads. (D) Deletions in chemical oligonucleotide synthesis. The mean percentages of consensus reads with deleted nucleotides are shown. Where multiple batches were used, means were calculated from all batches. n = 3 replicates per batch and oligonucleotide type. BS, BioSearch; EF, Eurofins Genomics; SA, Sigma-Aldrich.

		Sequence variant 1			Sequence variant 2		
Manufacturer	Purification	Batch 1	Batch 2	Batch 3	Batch 1	Batch 2	Batch 3
IDT	Desalted	x	x	x	x	x	x
IDT	PAGE	x	x	x	x	x	x
IDT	HPLC	x					
IDT	PAGE and HPLC	x					
IDT	Ultramer	x	x	x	x	x	x
IDT	gBlocks	x			x		
Eurofins	Desalted	x					
Eurofins	PAGE	x					
Eurofins	Extremer	x					
Sigma-Aldrich	Desalted	x	x	x	x	x	x
Sigma-Aldrich	PAGE	x	x	x	x	x	x
Sigma-Aldrich	HPLC	x					
BioSearch	Desalted	x					
BioSearch	Dual HPLC	x					

Table 1. Overview of synthetic oligonucleotides analyzed.

Abbreviation: HPLC, high-performance liquid chromatography.

Results

EXPERIMENTAL DESIGN TO QUANTIFY ERRORS IN CHEMICAL OLIGONUCLEOTIDE SYNTHESIS USING DIGITAL SEQUENCING

To quantify errors in chemical oligonucleotide syntheses, we designed a synthetic oligonucleotide containing a 70 nucleotides-long part of the *TP53* gene, of which 40 nucleotides are target primer sequences. An artificial 32 nucleotides-long insert of all 16 possible dinucleotide combinations was introduced 5' to the reverse primer sequence (sequence variant 1) (Fig. 1, A). To evaluate the sequence context, we also designed a second oligonucleotide (sequence variant 2) where the order of dinucleotide sequences and the genomic *TP53* sequence between the target primer sequences was switched and used digital sequencing to analyze different oligonucleotide types (Fig. 1, B). The hairpin barcoding primers used incorporate a random 12 nucleotides-long UMI that is added to target molecules during an initial PCR step. Barcoded products are then amplified using universal Illumina primers in a second PCR step. Specific PCR products are purified using magnetic beads and finally sequenced. The barcode information is then used to correct for polymerase-induced errors by forming consensus sequences of reads containing the same UMI. At least 10 reads were required for each UMI to build a consensus sequence that is used in downstream analyses. The

sequencing assay, targeting *TP53*, was designed and tested as previously described (22) (Supplemental Fig. 1). We expect 1 UMI for each single-stranded template molecule and 2 UMIs per double-stranded template due to the UMI-incorporation strategy (22). All libraries were sequenced to similar depths, enabling similar error correction for all oligonucleotide types (Supplemental Fig. 2).

We purchased 14 different types of oligonucleotides from 4 manufacturers with various purity grades using sequence variant 1 (Table 1). All types are single-stranded oligonucleotides, except IDT gBlocks, which are double-stranded. Five oligonucleotide types were also purchased in 3 independent batches, where each batch was purchased with at least 2 weeks between orders. Digital sequencing removes errors incurred during library preparation and sequencing and Figure 1, C shows the total error rates per nucleotide for all oligonucleotide types with and without UMI error correction. To determine the background noise level of our sequencing assay, we also analyzed genomic DNA purified from the human breast cancer cell line MCF-7.

OLIGONUCLEOTIDE ERRORS ARE PREDOMINANTLY DELETIONS AND DEPEND ON BOTH MANUFACTURER AND PURITY

Truncations are expected in chemical oligonucleotide synthesis due to failure in binding a new

phosphoramidite monomer to the elongating nucleotide chain (20). There are 2 subtypes of truncated oligonucleotides. Molecules that are successfully capped after failed coupling will not participate in further synthesis. The alternative is that molecules are only temporarily blocked from participating in synthesis for 1 or a few cycles, resulting in a deletion.

The number of truncated oligonucleotides varied between manufacturers and purification approaches, with a single missing nucleotide being the most common type of error for all oligonucleotide types (Fig. 1, D). IDT gBlocks displayed fewest truncations with 99.3% full-length sequences, while BioSearch desalted oligonucleotides contained the most truncated sequences where only 86.5% of the molecules were complete. As expected, desalted oligonucleotides contained fewer full-length molecules than purified variants, except for Sigma-Aldrich primers, where desalted oligonucleotides outperformed the purified variants. After IDT gBlocks, IDT Ultramers (98.47% full-length molecules) and Eurofins PAGE (98.36%) performed best. For comparison, we observed no deletions in genomic DNA purified from the breast cancer cell line MCF-7.

Oligonucleotide synthesis errors also depend on sequence context. Strikingly, deletion frequencies increased toward the 5'-end of all oligonucleotide types, except for IDT gBlocks (Fig. 2, A). We also observed individual nucleotide positions with increased frequencies of deletion errors, such as in the IDT PAGE-purified oligonucleotide. The nucleotide position of deletions was significantly correlated between most oligonucleotide types, even for those with overall few deletions (Fig. 2, B). The mean deletion error per nucleotide across all oligonucleotide types was 0.176%, with IDT gBlocks displaying the lowest mean deletion frequency per nucleotide (0.019%) and Eurofins desalted the highest (0.598%; Supplemental Fig. 3). In molecules with 2 or more deletions, these tended to occur sequentially. In fragments with 2 deleted nucleotides, these were together in 63% of all molecules, whereas molecules with 3 or 4 deleted nucleotides displayed at least 2 consecutively deleted nucleotides in 87% of all molecules (Supplemental Fig. 4).

Interestingly, we also observed substitutions in synthetic oligonucleotide synthesis. The mean substitution error frequency across all oligonucleotide types and nucleotide positions was 0.025%, 7 times lower levels than the level of deletions. IDT desalted displayed the lowest mean number of substitutions per nucleotide (0.008%), whereas Eurofins Extremer showed the highest (0.064%) (Supplemental Fig. 3). The single largest substitution error (1.143%) for an individual nucleotide occurred in the Eurofins Extremer oligonucleotide (Fig. 2, A). The nucleotide position for substitution errors were also correlated between oligonucleotide types, although

the mean correlation coefficient was slightly lower for substitutions ($\rho_{mean} = 0.46$) compared to deletions ($\rho_{mean} = 0.54$) (Fig. 2, B). The reference cell-line DNA contained substitutions with a mean error frequency of 0.0005% (Fig. 2, A; Supplemental Fig. 3), which was 50 times below the mean of all oligonucleotide types (1-sample t test, $P < 0.001$). For all oligonucleotide types, the frequency of deletions was higher than that of substitutions, except for IDT gBlocks with 0.019% deletions compared to 0.021% substitutions (Supplemental Fig. 3). Furthermore, the nucleotide positions for deletion and substitution errors across all oligonucleotides were also weakly, but significantly, correlated with each other ($\rho = 0.27$, $P < 0.05$).

OLIGONUCLEOTIDE ERRORS ARE AFFECTED BY BATCH EFFECTS AND SEQUENCE CONTEXT

The observation that deletions and substitutions are highly correlated with nucleotide position implies that the sequence context is important in chemical synthesis. Therefore, we purchased 2 additional batches, several weeks apart from each other, for a subset of the oligonucleotide types (Table 1). For the new batches we observed the same trend of increased error frequency toward the 5'-end of the oligonucleotides (Fig. 3, A). We also noted that the aberrantly high errors for individual nucleotide positions were specific to individual batches. As for batch 1, deletions were more common than substitutions in the new batches, but substitutions were still above the background level of genomic DNA (Fig. 3, A; Supplemental Fig. 5). Similarly, both deletions and substitutions were significantly correlated with nucleotide position for batches 2 and 3, comparable to the correlations found for batch 1 (Supplemental Fig. 6). The mean total error (deletions and substitutions) for all nucleotide positions was significantly different between batches 1 and 3 as well as between batches 2 and 3, but not between batches 1 and 2 (2-way ANOVA, adjusted P-value < 0.05). The total error was significantly lower for IDT PAGE compared to IDT desalted variants across the three batches, but we observed no significant difference between Sigma-Aldrich PAGE and Sigma-Aldrich desalted (2-way ANOVA, adjusted P-value < 0.05). Across all batches, IDT PAGE displayed the lowest mean total error (0.09%), whereas IDT desalted showed the highest (0.21%).

We also analyzed error rates in another sequence context (sequence variant 2). When analyzing sequence variant 2, we observed a weaker bias for deletion errors toward the 5'-end of all oligonucleotide types, although deletions were still more common than substitutions (Fig. 3, A; Supplemental Fig. 7). Correspondingly, the nucleotide positions for deletions were not significantly correlated between most oligonucleotide types (Supplemental Fig. 8). However, substitution errors
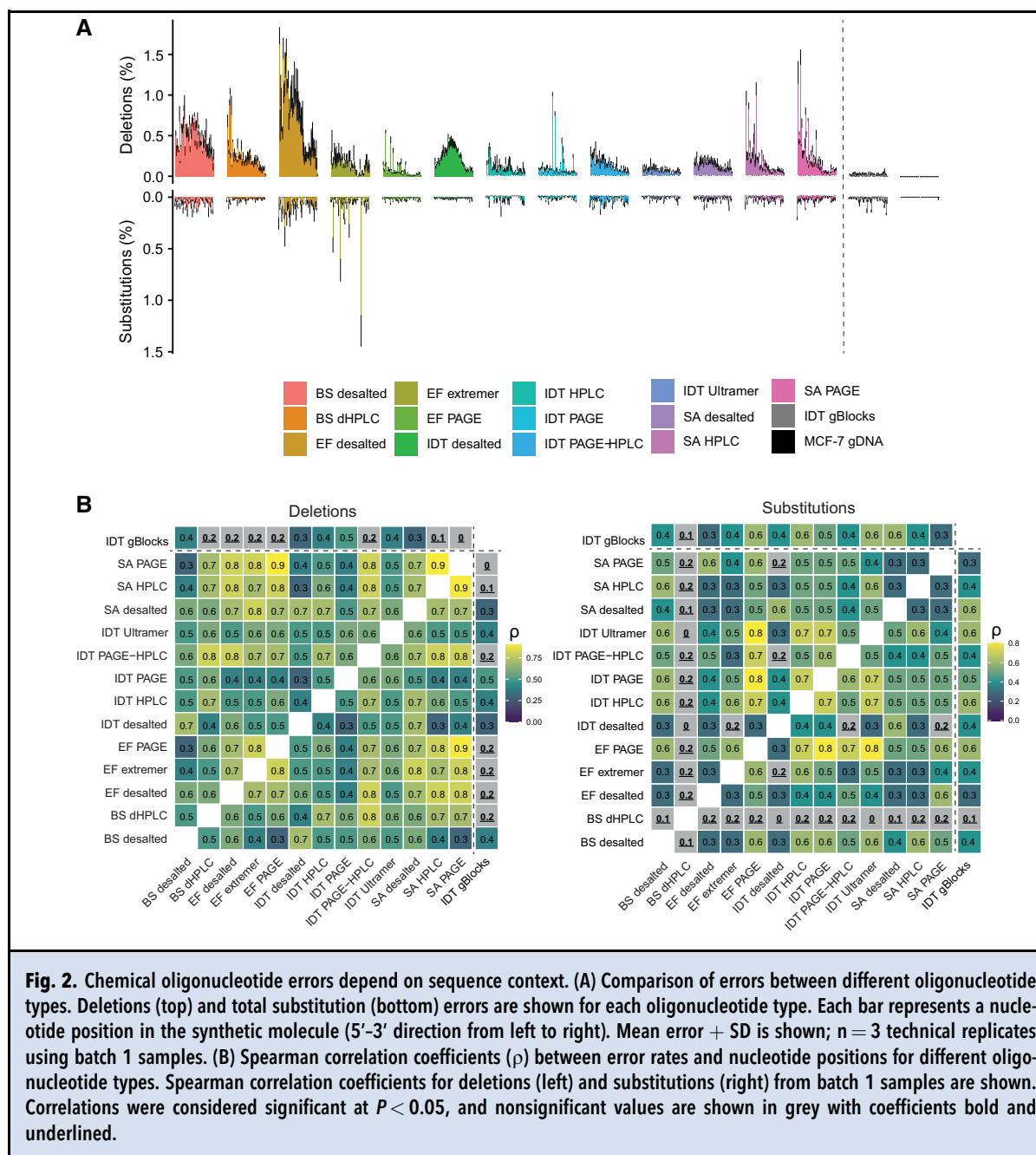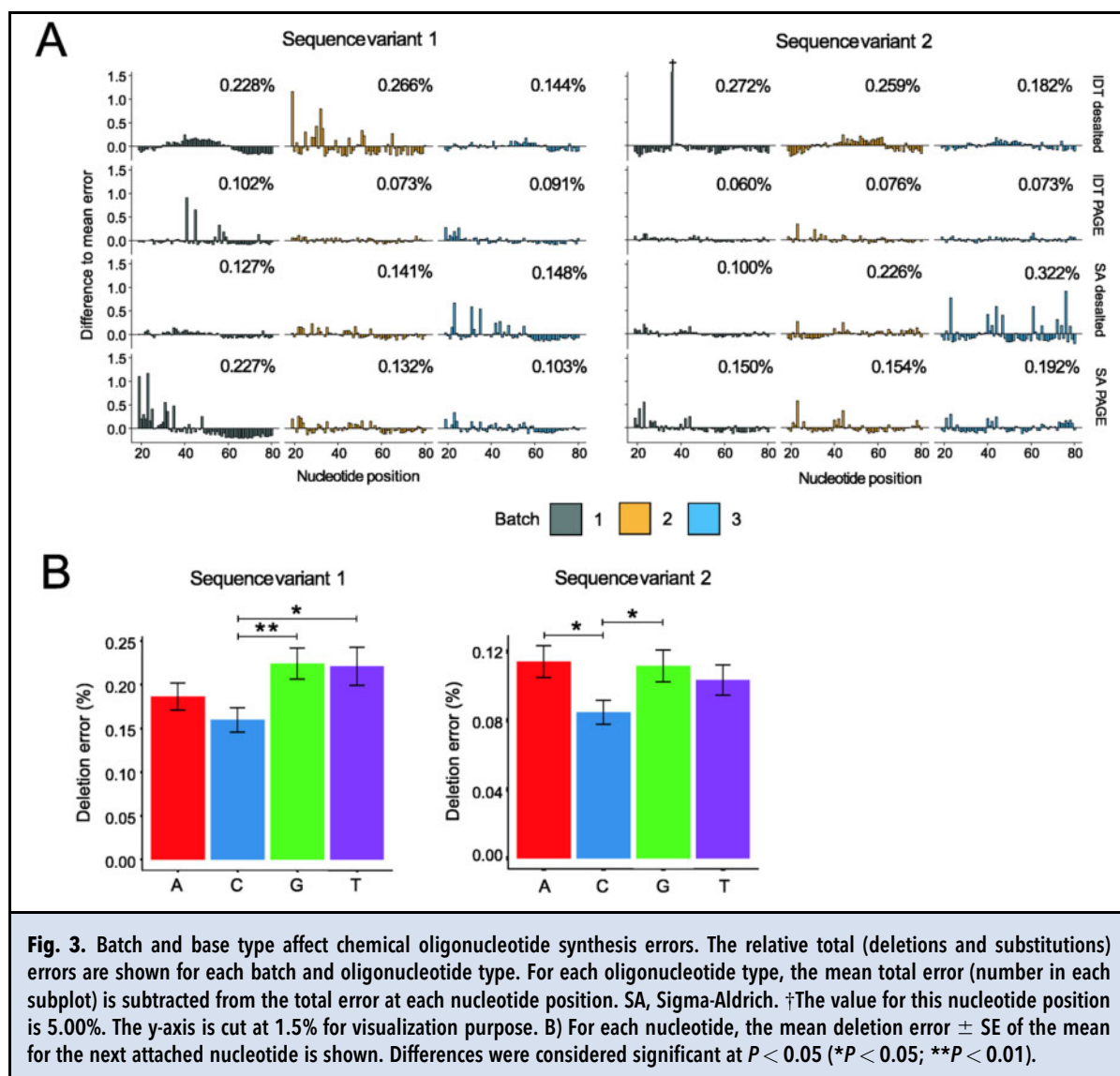
Fig. 2. Chemical oligonucleotide errors depend on sequence context. (A) Comparison of errors between different oligonucleotide types. Deletions (top) and total substitution (bottom) errors are shown for each oligonucleotide type. Each bar represents a nucleotide position in the synthetic molecule (5'-3' direction from left to right). Mean error $+$ SD is shown; n $=$ 3 technical replicates using batch 1 samples. (B) Spearman correlation coefficients (ρ) between error rates and nucleotide positions for different oligonucleotide types. Spearman correlation coefficients for deletions (left) and substitutions (right) from batch 1 samples are shown. Correlations were considered significant at $P < 0.05$, and nonsignificant values are shown in grey with coefficients bold and underlined.

were still significantly correlated with nucleotide positions between most oligonucleotide types. We again detected batch-specific errors for individual nucleotide positions (Fig. 3, A). The mean total errors were significantly different between batches 1 and 3, but not between any other batch comparisons (2-way ANOVA, adjusted P-value < 0.05). The difference in total error was again significantly lower for IDT PAGE compared to IDT desalted variants across the 3 batches, whereas

the difference between Sigma-Aldrich PAGE and Sigma-Aldrich desalted was not significant (2-way ANOVA, adjusted P-value < 0.05). The mean total error was again lowest for IDT PAGE (0.07%) and highest for IDT desalted (0.24%).

We also analyzed the 16 dinucleotide sequence combinations separately and found that the preceding nucleotide affects the error rate of the newly attached base. For both sequence variants the errors were
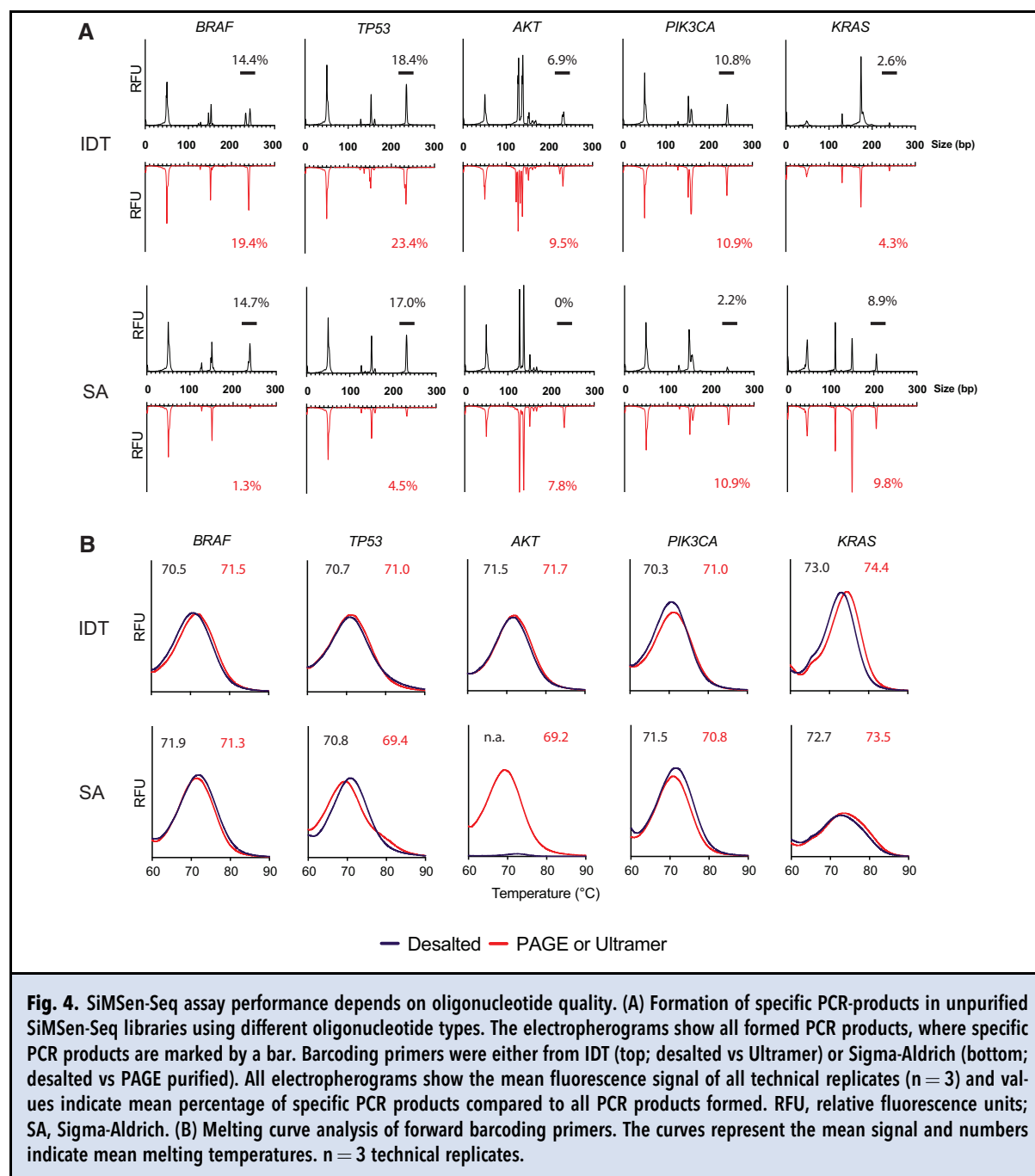
Fig. 3. Batch and base type affect chemical oligonucleotide synthesis errors. The relative total (deletions and substitutions) errors are shown for each batch and oligonucleotide type. For each oligonucleotide type, the mean total error (number in each subplot) is subtracted from the total error at each nucleotide position. SA, Sigma-Aldrich. †The value for this nucleotide position is 5.00%. The y-axis is cut at 1.5% for visualization purpose. B) For each nucleotide, the mean deletion error ± SE of the mean for the next attached nucleotide is shown. Differences were considered significant at $P < 0.05$ (*$P < 0.05$; **$P < 0.01$).

significantly lower if the new base was attached to a cytosine (Fig. 3, B; Supplemental Fig. 9, A). We also observed 2.6 and 2.3 times increase in deletion error rates if the same base type was added after each other for sequence variant 1 and 2, respectively (*t* test, $P < 0.05$; Supplemental Fig. 9, B).

OLIGONUCLEOTIDE QUALITY IS CRITICAL FOR ASSAY PERFORMANCE

High-quality oligonucleotides are critical components in several molecular techniques, including SiMSeq-Seq assays. Here, we tested 5 different SiMSen-Seq assays targeting sequences in *AKT*, *BRAF*, *KRAS*, *PIK3CA*,

and *TP53* using either desalted or purified barcoding primers (Fig. 1, B) from IDT or Sigma-Aldrich. The assays were evaluated by analyzing their capacity to form specific PCR products for sequencing using capillary gel electrophoresis and by melting curve analysis of the stem-loop structure in the forward SiMSen-Seq barcoding primer (Fig. 4, A and B). For the *TP53* and *BRAF* assays, we observed that PAGE-purified barcoding primers from Sigma-Aldrich contained less specific PCR product in unpurified libraries than the desalted barcoding primer variants, whereas the Ultramer version from IDT generated more specific PCR products compared to the desalted barcoding primer variants (Fig. 4, A).

Fig. 4. SiMSen-Seq assay performance depends on oligonucleotide quality. (A) Formation of specific PCR-products in unpurified SiMSen-Seq libraries using different oligonucleotide types. The electropherograms show all formed PCR products, where specific PCR products are marked by a bar. Barcoding primers were either from IDT (top; desalted vs Ultramer) or Sigma-Aldrich (bottom; desalted vs PAGE purified). All electropherograms show the mean fluorescence signal of all technical replicates (n = 3) and values indicate mean percentage of specific PCR products compared to all PCR products formed. RFU, relative fluorescence units; SA, Sigma-Aldrich. (B) Melting curve analysis of forward barcoding primers. The curves represent the mean signal and numbers indicate mean melting temperatures. n = 3 technical replicates.

These data matched with the melting temperature of the stem-loop structure, where the worse performing assays, PAGE for Sigma-Aldrich and desalted for IDT, displayed lower melting temperatures, indicating more chemical synthesis errors in the stem-loop sequences (Fig. 4, B). For the *AKT*, *PIK3CA*, and *KRAS* assays, the amount of specific PCR-products increased for the purified oligonucleotide types. This was especially

pronounced for the *AKT* assay that failed to generate any specific PCR-products using the Sigma-Aldrich desalted barcoding primers, while the assay was successful with PAGE purified barcoding primers. This was also reflected in the melting temperature analysis, where the fluorescence signal for the Sigma-Aldrich desalted forward barcoding primer was barely detectable. For the *PIK3CA* and *KRAS* assays, melting temperatures also

matched the amount of specific PCR products, except for the *PIK3CA* assay using Sigma-Aldrich primers (Fig. 4).

Discussion

Oligonucleotides are critical components of many molecular techniques, including ultrasensitive mutation detection, cloning, and single-cell analysis. For instance, randomized nucleotide sequences known as molecular barcodes are used for noise reduction and sample deconvolution, and errors in barcodes can become a source of bias (23). Numerous manufacturers offer a broad range of oligonucleotide types and purifications, yet their performance is often difficult to quantify. Here, we used digital sequencing to comprehensively analyze the error profiles in chemically synthesized oligonucleotides across multiple combinations of manufacturers, purity grades, batches, and sequence variants. A limitation of our PCR-based sequencing approach is that we can only access the oligonucleotides if both primers can bind and amplify target DNA during the barcoding PCR step. The use of UMIs will correct for uneven amplification, but 5'-truncated oligonucleotide molecules cannot be amplified, and molecules with several errors in the primer regions will, most likely, not be amplified. We speculate that this is more likely for lower purity grades, such as desalted oligonucleotide types. We used comparable numbers of amplifiable molecules in library construction. Hence, our data are somewhat biased in favor of desalted oligonucleotides since we adjusted for nonamplifiable molecules. Despite this bias, overall lower purity grades performed worse than higher grade oligonucleotide types.

A previous study showed that up to 17% of all oligonucleotide molecules contained single deletion errors mostly at the 3'-end of a 25 nucleotides-long molecule, while this type of error decreased considerably toward the 5'-end (19). Nonetheless, virtually all nucleotide positions contained deletions. Another report showed a uniform profile of deletions across a 21 nucleotides long molecule with no dependence on either base type or nucleotide position (20). Except for severely truncated molecules that are missing their 5'-ends, we show that chemical synthesis errors are predominantly deletions, although substitution errors are also present at virtually every position. Except for IDT gBlocks, deletions were more common than substitutions (Fig. 2, A). It should be noted that gBlocks are different from the other conventional oligonucleotide types as they are double-stranded and must be at least 125 nucleotides long. Being double-stranded could also explain why no directional bias from 3' to 5' was observed for gBlocks as synthesis may occur in both strands.

The predominant type of deletions are single missing bases, although fragments containing multiple deletions exist even for the highest purification levels. Deletions presumably occur when capping fails to prevent oligonucleotides with failed coupling to participate in downstream coupling steps. However, this does not explain why most deletions occur consecutively in oligonucleotides with several deletions. We speculate that some oligonucleotides are temporarily blocked physically from participating in the coupling and capping steps for several cycles of synthesis. The observation of substitution errors suggests that nucleotides are occasionally incorporated wrongly during chemical oligonucleotide synthesis. The exact cause of substitution errors is unclear, but we speculate that a potential reason could be incomplete removal of specific bases between each synthesis step.

Our data suggest that some oligonucleotide sequences show a directional bias in chemical oligonucleotide synthesis errors with deletion frequencies increasing in the direction of synthesis. We also found that some batches of oligonucleotides displayed a disproportional number of deletions at individual nucleotide positions. These errors were at relatively high frequencies (about 1%), even for PAGE-purified oligonucleotides that otherwise displayed low error rates. Additionally, the difference in total errors between some batches of Sigma-Aldrich desalted oligonucleotides was 3.2 times, but the mean reduction in errors through PAGE purification of the same sequence variant and manufacturer was only 1.3 times (Fig. 3, A). These data show that oligonucleotide batch can be more important than choice of purification method to obtain oligonucleotides with minimal number of errors. The underlying reasons for the observed batch variations are unclear but may depend on oligonucleotide synthesizer, operator, and reagents.

Oligonucleotide quality directly impacts SiMSen-Seq library yield. We observed up to 11 times increase in specific PCR product formation for the same assay by choosing a specific oligonucleotide type as barcoding primers. Manufacturers often recommend PAGE purification for oligonucleotides >50 to 60 bases, such as SiMSen-Seq primers, as it is supposed to generate the highest level of purity, albeit with reduced yield compared to other methods. However, we find that oligonucleotide purification does not always improve assay performance. Some PAGE-purified assays from Sigma-Aldrich produced worse libraries than desalted barcoding primer variants for the same assays. These findings agree with our data showing that Sigma-Aldrich PAGE-purified oligonucleotides contained more deletion errors compared to the desalted alternatives from the same manufacturer (Fig. 3, A). We show that integrity of

Fig. 5. Chemical synthesis errors affect assay performance. (A) Applications that rely on secondary structures, like SiMSen-Seq, may be adversely affected by oligonucleotide errors. (B) Deletions and substitutions in primers may generate complementary sequences that increase the formation of unspecific PCR products. (C) Errors in primers may result in reduced ability to bind target DNA efficiently, especially in application using low primer concentrations, such as preamplification and single cell analysis. (D) Erroneous primer sequences may result in off-target amplification. (E) Minor errors in primer sequences may result in amplicon switching in multiplex PCR in which similar sequences are targeted. For example, errors in degenerate primers may cause amplicon switching.

the stem-loop structure correlates with library quality, which further strengthens the link between errors in the oligonucleotide sequence and assay performance. A potential strategy to improve oligonucleotide quality is to apply enzymatic error correction, an affordable technique for removal of synthesis errors (24–26). Further studies are needed to determine the systematic effects of various purification methods.

Chemically synthesized oligonucleotides with errors may adversely affect several applications, such as PCR, digital PCR, and sequencing (Fig. 5). For example, some applications, such as SiMSen-Seq, rely on the formation of secondary DNA structures, which may be destabilized through errors in barcoding primers. In numerous applications, erroneous primer sequences may result in increased formation of primer dimers, reduced target DNA/RNA binding, and unspecific target identification. These limitations are especially problematic in approaches that require single molecule resolution.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

References

1. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 1986;51(Pt 1):263–73.

2. Gilboa E, Mitra SW, Goff S, Baltimore D. A detailed model of reverse transcription and tests of crucial aspects. Cell 1979;18: 93–100.

3. Seo TS, Bai X, Kim DH, Meng Q, Shi S, Ruparel H, et al. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. Proc Natl Acad Sci USA 2005;102:5926–31.

4. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res 2013;41: 4336–43.

5. Smith HO, Hutchison CA 3rd, Pfannkoch C, Venter JC. Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. Proc Natl Acad Sci USA 2003; 100:15440–5.

6. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell 2016;167:1853–66.

7. Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, et al. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res 1989;17:2503–16.

8. Petersen DW, Kawasaki ES. Manufacturing of microarrays. Adv Exp Med Biol 2007;593:1–11.

9. Smith AJ, Humphries SE. Characterization of DNA-binding proteins using multiplexed competitor EMSA. J Mol Biol 2009;385:714–7.

10. Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. Nat Rev Genet 2012;13: 358–69.

11. Ma S, Saaem I, Tian J. Error correction in gene synthesis technology. Trends Biotechnol 2012;30:147–54.

12. Caruthers MH, Beaucage SL, Becker C, Efcavitch JW, Fisher EF, Galluppi G, et al. Deoxyoligonucleotide synthesis via the phosphoramidite method. Gene Amplif Anal 1983;3:1–26.

13. Sastry SS, Spielmann HP, Dwyer TJ, Wemmer DE, Hearst JE. Recent advances in the synthesis and structure determination of site specifically psoralen-modified DNA oligonucleotides. J Photochem Photobiol B 1992;14: 65–79.

14. Pazdernik N,S. Getting enough full-length oligo? Coupling efficiency matters! https://eu.idtdna.com/pages/education/decoded/article/getting-enough-full-length-oligo (Accessed April 2021).

15. Merck. DNA oligonucleotide synthesis. https://www.sigmaaldrich.com/technical-documents/articles/biology/dna-oligonucleotide-synthesis.html (Accessed April 2021).

16. Gilar M. Analysis and purification of synthetic oligonucleotides by reversed-phase high-performance liquid chromatography with photodiode array and mass spectrometry detection. Anal Biochem 2001;298:196–206.

17. Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, Church G. Accurate multiplex gene synthesis from programmable DNA microchips. Nature 2004;432:1050–4.

18. Borovkov AY, Loskutov AV, Robida MD, Day KM, Cano JA, Le Olson T, et al. High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. Nucleic Acids Res 2010;38:e180.

19. Temsamani J, Kubert M, Agrawal S. Sequence identity of the n-1 product of a synthetic oligonucleotide. Nucleic Acids Res 1995;23:1841–4.

20. Chen D, Yan Z, Cole DL, Srivatsa GS. Analysis of internal (n-1)mer deletion sequences in synthetic oligodeoxyribonucleotides by hybridization to an immobilized probe array. Nucleic Acids Res 1999;27: 389–95.

21. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 2012;13:134.

22. Ståhlberg A, Krzyzanowski PM, Egyud M, Filges S, Stein L, Godfrey TE. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. Nat Protoc 2017;12:664–82.

23. Ståhlberg A, Krzyzanowski PM, Jackson JB, Egyud M, Stein L, Godfrey TE. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. Nucleic Acids Res 2016;44:e105.

24. Lee DF, Lu J, Chang S, Loparo JJ, Xie XS. Mapping DNA polymerase errors by single-molecule sequencing. Nucleic Acids Res 2016;44:e118.

25. Matzas M, Stahler PF, Kefer N, Siebelt N, Boisguerin V, Leonard JT, et al. High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. Nat Biotechnol 2010;28: 1291–4.

26. Lubock NB, Zhang D, Sidore AM, Church GM, Kosuri S. A systematic comparison of error correction enzymes by next-generation sequencing. Nucleic Acids Res 2017;45: 9206–17.